統計数理(2012)
第 60 巻 第 2 号 305–316

特集「多様性と進化の統計解析」
[研究詳解]

ウイルスタンパク質変異にかかる多様化圧の空間分布

渡部 輝明[1] ・岸野 洋久[2]

(受付 2011 年 12 月 27 日；改訂 2012 年 5 月 23 日；採択 6 月 6 日)

要　旨

タンパク質は従来から備えている機能を変化させたり，新規に機能を獲得したりすることで環境に適応する能力を有している．この環境適応はアミノ酸配列置換によってもたらされるが，アミノ酸配列の置換は遺伝子上で起こった突然変異が淘汰された結果である．そのため突然変異の淘汰は環境に特有の選択圧のもとで行われ，進化の過程において変化する環境に依存していくものと考えられる．つまりタンパク質適応進化について理解するためには，選択圧の時空間的な揺らぎを解明することが重要となる．我々は階層ベイズモデルを通して，タンパク質表面における選択圧の空間分布を検出する方法を開発した．この方法は磁性体物理の分野で理論的枠組みが確立したイジング模型を用いて構築されており，選択圧の空間集積性の強さと広さを決める超パラメータは周辺尤度を最大化することで決定可能である．事前分布のモデルは正規化が困難であるため，熱力学的積分のアイデアを利用した方法により対数周辺尤度を計算する．この方法をインフルエンザウイルスのヘマグルチニンタンパク質に適用し，選択圧の空間分布を検出した．

キーワード： 分子進化，選択圧，空間分布，階層ベイズモデル，イジング模型．

1. 遺伝子変異にかかる選択圧

遺伝子の塩基配列データに基づく種間の系統関係推定は，分子進化速度が一定であることを仮定して行われることが多い．これは分子時計が成り立つとした分子進化の中立説(Kimura, 1983)に根拠を持つ．しかし多くの場合，進化速度は変動し，その要因は様々である．タンパク質のアミノ酸配列を決定するコード領域においては，同義置換速度と非同義置換速度の変動として進化速度の変動を捉えることが出来る．同義置換(synonymous substitution)とはアミノ酸を変えない塩基置換を指し，非同義置換(non-synonymous substitution)とはアミノ酸を変える塩基置換を意味する．3 つの塩基の並びで構成されるコドンの一例として “AAA” はリジン(Lys)をコードしているが，3 番目の塩基がアデニン(A)からグアニン(G)に置換し “AAG” となってもコードするアミノ酸はリジンのまま変わらない(同義置換)．これに対して，3 番目の塩基がアデニンからシトシン(C)に置換し “AAC” となるとアスパラギン(Asn)をコードするようになる(非同義置換)．同義置換はアミノ酸を変えないため自然淘汰の影響はあまり受けないと考えられる．一方，非同義置換は該当するアミノ酸残基がタンパク質の内部に位置する場合，

[1] 高知大学 医学部附属医学情報センター：〒783–8505 高知県南国市岡豊町小蓮
[2] 東京大学 農学生命科学研究科：〒113–8657 東京都文京区弥生 1–1–1

構造安定性維持の理由から淘汰される可能性が高い．また，タンパク質表面に位置するアミノ酸を変えるような置換では，タンパク質機能維持の理由からこれも淘汰される場合が多いと考えられる．これらのような状況では非同義置換速度は同義置換速度よりも遅くなる．

非同義置換は自然淘汰の結果，同義置換よりも排除される場合が多いと先に述べたが，アミノ酸変異がタンパク質の適応度において優位なものである場合にはその限りではない．本稿で扱うインフルエンザウイルスのヘマグルチニンタンパク質(HA)は宿主細胞の受容体へ結合し，ウイルス粒子が宿主細胞へ侵入する最初の足がかりを形成する重要なタンパク質であるが，ウイルス粒子の膜上に存在するため宿主免疫系からの攻撃にさらされる．抗体(免疫グロブリン)が HA タンパク質の宿主細胞受容体結合領域へ結合し，ウイルス粒子の宿主細胞への侵入を妨げるのである．そのため HA タンパク質の受容体結合領域では抗体の結合能を下げるアミノ酸変異が早く定着するが，同時に受容体との結合能を下げてしまう恐れがある．受容体との結合能を維持する必要から抗体の脅威が軽減されると受容体との結合能を回復するアミノ酸変異が定着していくようである(Watabe et al., 2007)．また，ウイルス粒子の受容体結合能の維持と抗体結合からの回避のバランスから定義される適応度地形においては，固定確率の非常に高い(>0.9)置換が起きうるアミノ酸残基が存在する(Watabe and Kishino, 2010)．

このように非同義置換速度はタンパク質構造における空間的な位置によって変動することが考えられる．本稿では我々の最近の研究(Watabe and Kishino, 2012)から，コドンモデルを用いた選択圧の検出をタンパク質構造情報に根ざして行う手法について解説する．

2. コドンモデル

コドンモデルは 3 つの塩基の並びで構成されるコドンの置換を表現したものである．DNA の各塩基座位では 4 種類の塩基(A，T，G，C)が可能であるため，コドン総数はアミノ酸を指定しない 3 つの終止コドン(TAG，TAA，TGA：核遺伝子の場合)を除いて 61 種類($=4^3-3$)ある．我々の開発した方法ではアミノ酸残基毎に非同義置換速度と同義置換速度の比($\omega = dN/dS$)が異なることを取り入れているため，k 番目のアミノ酸残基におけるコドン i からコドン j への突然変異の瞬間速度は以下のように表現される：

$$q_{ij}^{(k)} = \begin{cases} 0 & \text{for more than one nucleotide substitution between } i \text{ and } j \\ u^{(k)}\pi_j & \text{for synonymous transversion} \\ u^{(k)}\kappa\pi_j & \text{for synonymous transition} \\ u^{(k)}\omega^{(k)}\pi_j & \text{for nonsynonymous transversion} \\ u^{(k)}\omega^{(k)}\kappa\pi_j & \text{for nonsynonymous transition.} \end{cases} \tag{2.1}$$

ここで π_j はコドン j の頻度を表し，解析対象とする塩基配列データから求められる．κ は転位型(transition)の塩基置換(A $\leftrightarrow$ G と T $\leftrightarrow$ C)の速度と転換型(transversion)の塩基置換(転位型以外)の速度の比を表す．同一コドン間変異の瞬間速度は $q_{ii}^{(k)} = -\sum_{j\neq i} q_{ij}^{(k)}$ で与えられる．そして $u^{(k)}$ は単位時間に起こるコドン置換を標準化($-\sum_{i=1}^{61}\pi_i q_{ii}^{(k)} = 1$)するように決められる．

k 番目のコドン座位におけるコドン i からコドン j への遷移確率は置換数 t に対して $p_{ij}^{(k)}(t) = (\exp\{\boldsymbol{q}^{(k)}t\})_{ij}$ で与えられる．ここで太字表示の $\boldsymbol{q}^{(k)}$ は(2.1)式の行列表現である．各コドン座位の尤度は，系統樹を構成する枝にわたり隣接節間の状態の推移を $p_{ij}^{(k)}(t)$ で記述し，潜在変数である内部節の状態について足し合わせをすることにより計算することが出来る(Felsenstein, 1981)．

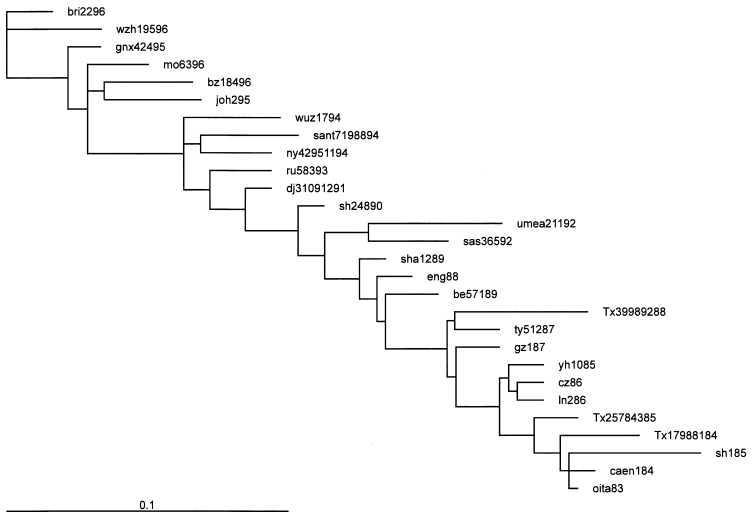

図 1. A 型香港風邪の Flu-HA タンパク質の約 40 年に渡る進化過程を表した系統樹.

3. インフルエンザウイルスの HA タンパク質の分子進化系統

タンパク質の配列変異は環境に応じた選択圧のもとで淘汰され，その選択圧はタンパク質の空間的な(機能を担う)部位のみならず進化過程での時間的な(機能を獲得した)時点においても変動する．これまで連続的な抗原性の変異などが配列データから詳細に解析されてきた(Cox and Bender, 1995; Smith et al., 2004)．現在では数千株規模の配列データが解析可能であるが(Bao et al., 2008)，本稿ではそのごく一部の 30 株弱を解析する．図 1 は A 型香港風邪のインフルエンザウイルス HA タンパク質の 1984 年から 1996 年に渡る進化過程を表した系統樹である(Yang et al., 2000)．非同義置換速度と同義置換速度の比 (ω) は各コドン座位(アミノ酸残基)で共通であるとして系統樹を得ている．この進化系統樹のどの枝(時間的位置)でどの空間的部位の選択圧が強く(又は弱く)なっているかを検出することで，宿主免疫系との対峙の仕方が見て取れる．特に幹を成す経路で選択圧が強く出るのか，末梢の枝で起こることなのかを知ることは，図 1 に観られる幹と末梢の系統樹構造を示すウイルス進化の理解を大きく進めることにつながる．本稿では選択圧の空間的な揺らぎに焦点をあてて解析を進める．

4. タンパク質表面での選択圧の揺らぎ

非同義置換の速度が同義置換に比べて大きい場合 ($\omega>1$)，選択圧は多様化をもたらすように働き，逆に小さい場合 ($\omega<1$) は変異を浄化するように働く．この選択圧の解析はこれまでほとんどの場合，配列情報のみによってされてきており(Suzuki and Gojobori, 1999; Suzuki, 2004a; Yang et al., 2000)，情報量の不足から偽陽性が多く検出されたり，タンパク質立体構造とは整合性が保証されていない選択圧分布が配列上に得られたりしていた．立体構造情報を解析に取り入れて情報量の増加を図り，同時に立体構造と整合性のとれた選択圧分布(立体構造における選択圧の空間分布)を得るために Suzuki (2004b) は three-dimensional window analysis を開発した．最尤法の枠組みでは，この方法は局所尤度法(Tibshirani and Hastie, 1987)として捉えることができ，配列情報に加えて立体構造情報を取り入れた優れた方法である．しかし球形領

域で定義される window の半径をデータから推定することが出来ず，結果に任意性を残してしまう．我々はイジング模型を事前分布に用いる階層ベイズモデルを開発し，この問題を解決した．選択圧の空間集積性の強さと広さを決める超パラメータは周辺尤度を最大化することにより，全てデータから推定出来る．イジング模型は画像修復の領域(Inoue and Tanaka, 2001)や森林育成のシミュレーション(Schlicht and Iwasa, 2004)などに応用されている模型であり，その理論的な枠組みは磁性体物理学の分野で確立され，パラメータ空間に分布し互いに作用し合う要素の熱力学的挙動を調べることを可能にする模型である．

4.1 周辺尤度

遺伝子配列情報とタンパク質立体構造情報及び配列間系統関係が与えられたときの “モデル” の尤度は周辺尤度で与えられる：

$$Z = P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{X},\boldsymbol{T},M).$$

モデルはタンパク質立体構造情報及び配列間系統関係とは独立であるとし，立体構造情報と配列間系統関係は所与としている．ここで $\boldsymbol{A}^{(i)}$ は配列長が $3L$ の全部で N 配列ある i 番目の塩基配列を表している，$\boldsymbol{A}^{(i)} = (a_1^{(i)},\ldots,a_L^{(i)})$．$a_k^{(i)}$ は i 番目の塩基配列の k 番目のコドンを表している．$\boldsymbol{X}$ はアミノ酸残基の α 炭素原子の位置を表している，$\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L)$．ここで $\boldsymbol{x}_k$ は k 番目の残基の α 炭素原子の空間座標を表している，$\boldsymbol{x}_k = (x_k, y_k, z_k)$．これら α 炭素原子の空間座標はアミノ酸置換によって変化しないことを仮定する．$\boldsymbol{T}$ は配列間の系統関係を表す系統樹を示し，系統樹を構成する “枝” の集合を表している．本稿ではモデル M を通してタンパク質変異にかかる選択圧の空間分布を明らかにすることを想定していることから，以下の様に置換速度比を表すパラメータを明示的に導入する：

$$Z = \int P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{\omega},\boldsymbol{X},\boldsymbol{T},M)P(\boldsymbol{\omega}|\boldsymbol{X},\boldsymbol{T},M)d\boldsymbol{\omega}.$$

各アミノ酸残基における置換速度比はベクトル表記されている，$\boldsymbol{\omega} = (\omega_1,\ldots,\omega_L)$.

4.2 熱力学的積分

事前分布 $P(\boldsymbol{\omega}\,|\,\boldsymbol{X},\boldsymbol{T},M)$ が規格化されていない場合，周辺尤度の表記には事前分布のパラメータ ω_k による積分を分母に持つ必要がある：

$$Z = \frac{\int P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{\omega},\boldsymbol{X},\boldsymbol{T},M)P(\boldsymbol{\omega}\,|\,\boldsymbol{X},\boldsymbol{T},M)d\boldsymbol{\omega}}{\int P(\boldsymbol{\omega}'\,|\,\boldsymbol{X},\boldsymbol{T},M)d\boldsymbol{\omega}'}. \tag{4.1}$$

しかし一般的にパラメータの全空間に渡る積分は現実には困難を伴う．Ogata (1989)はこの問題を Monte Carlo 法を用いて一般的な枠組みで解決した．この方法は，熱力学的積分(Thermodynamic integration)としてその応用範囲を広げている(Gelman and Meng, 1998; Lartillot and Philippe, 2006)．

(4.1)式において周辺尤度の対数をとり，右辺が分子の対数と分母の対数の差であることを形式的にパラメータ β の導入により表現する：

$$\ln Z = \int_0^1 \frac{d}{d\beta}\left\{\ln \int P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{\omega},\boldsymbol{X},\boldsymbol{T},M)^{\beta} P(\boldsymbol{\omega}\,|\,\boldsymbol{X},\boldsymbol{T},M)d\boldsymbol{\omega}\right\}d\beta.$$

この後はパラメータ β による微分を遂行し，パラメータ ω_k の全空間に渡る積分を回避する形式を得る．パラメータ β により特徴付けられる系における期待値を計算することで周辺尤度の対数を得るのである：

$$\ln Z = \int_0^1 E_\beta[\ln P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{\omega},\boldsymbol{X},\boldsymbol{T},M)]d\beta, \tag{4.2}$$

ここで $E_\beta[\cdots]$ は以下の密度関数を考慮した期待値を意味する:

$$d_\beta(\boldsymbol{\omega}) = \frac{P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{\omega},\boldsymbol{X},\boldsymbol{T},M)^\beta P(\boldsymbol{\omega}\,|\,\boldsymbol{X},\boldsymbol{T},M)}{\int P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{\omega}',\boldsymbol{X},\boldsymbol{T},M)^\beta P(\boldsymbol{\omega}'\,|\,\boldsymbol{X},\boldsymbol{T},M)d\boldsymbol{\omega}'}.$$

4.3 イジング模型を用いたギブスサンプリング

期待値 $E_\beta[\cdots]$ を計算するためにギブスサンプリングを用いる．置換速度比の事前分布は ω_k を 3 つの状態 $(n=3)$ に階級化したイジング模型によりモデル化する:

$$P(\boldsymbol{\omega}\,|\,\boldsymbol{X},\boldsymbol{T},M) = \exp\left\{\lambda\sum_{l>k}(\boldsymbol{J}(\alpha,r_{kl}))_{s_k s_l}\right\}. \tag{4.3}$$

この模型は，格子上スピンの相互作用を記述するイジング模型を拡張し，3 次元空間に分布するスピン間の相互作用をそれらの空間距離を考慮して扱えるようにしたものである．ここで s_k は階級化された置換速度比を示す指標である．選択圧が変異を浄化するように働いていることを示す領域 $(\omega_k<1)$ で $s_k=1$，変異が中立的であることを示す領域 $(\omega_k\sim1)$ で $s_k=2$，そして選択圧が多様化をもたらすように働くことを示す領域 $(\omega_k>1)$ で $s_k=3$ をとる．タンパク質の 3 次元構造 $\boldsymbol{X}$ はアミノ酸残基の α 炭素原子間の空間距離 r_{kl} のみを用いる．また，置換速度比は系統樹上の位置には依らないことを仮定している．$\boldsymbol{J}$ の具体的な定義は以下で与えられる:

$$(\boldsymbol{J}(\alpha,r))_{s's} = (\exp\{\alpha r\boldsymbol{Q}\})_{s's} - \frac{1}{n}.$$

ここで $\boldsymbol{Q}$ は n 行 n 列行列であり，対角要素が $-(n-1)$ で非対角要素が 1 である．ここで n は ω_k の状態数 $(n=3)$ を表している．この $\boldsymbol{J}$ の定義では近距離にあるアミノ酸残基の組は遠距離にある組より高い相関を持つことが保証されている．図 2 に $\alpha=0.1,\ 0.3,\ 0.5$ の場合で r が 1.5 nm (15Å) までの $\boldsymbol{J}$ の具体的な様子を示した．α 炭素原子間の距離は最短でおよそ 0.4 nm (4Å) 程度なので $\alpha=0.5$ の場合ではほとんど事前分布（(4.3)式）が寄与しないことが判る．

アミノ酸配列の進化過程で各アミノ酸残基での置換は，タンパク質の 3 次元構造においてその残基が位置する部位の担う機能により影響を受けると考えられる．ここではそれらの影響を置換速度比としてのみ表現し，配列が与えられた時の ω_k の尤度を各残基での尤度の積として表す:

$$P(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(N)}\,|\,\boldsymbol{\omega},\boldsymbol{X},\boldsymbol{T},M) = \prod_{k=1}^{L} P(\boldsymbol{a}_k\,|\,\omega_k,\boldsymbol{T}), \tag{4.4}$$

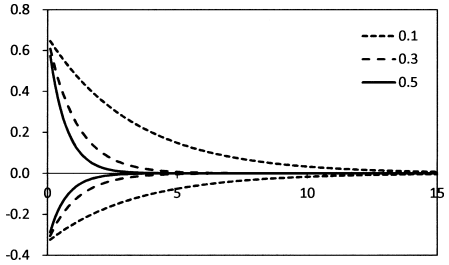

図 2. $\boldsymbol{J}$ の具体的な様子．対角要素では正の値をとり，非対角要素では負の値を持つ．横軸は α 炭素原子間の距離をÅ 単位で表示している．

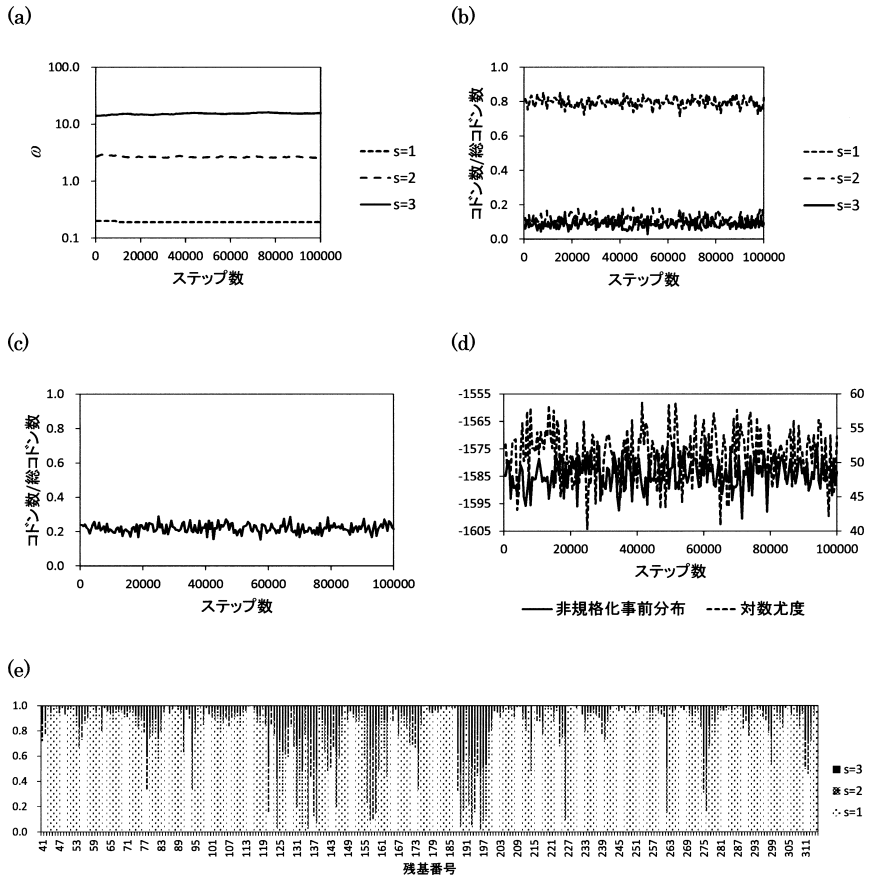

図 3. $\lambda=4.0$, $\alpha=0.165$, $\beta=1.0$ の場合. (a) ω の階級値の MCMC 軌跡. (b) 3 階級の組成の MCMC 軌跡. (c) 階級遷移率の MCMC 軌跡. (d) 非規格化事前分布と対数尤度の MCMC 軌跡. (e) 各残基における ω の 3 階級の事後確率.

ここで $\boldsymbol{a}_k$ は N 本のアミノ酸配列の k 番目のアミノ酸を要素とするベクトルである, $\boldsymbol{a}_k=(a_k^{(1)},\ldots,a_k^{(N)})$. 各残基における尤度 $P(\boldsymbol{a}_k|\omega_k,\boldsymbol{T})$ は系統樹 $\boldsymbol{T}$ が与えられている下で, コドン間(コドン i から置換数 t を経てコドン j へ)の遷移確率 $p_{ij}^{(k)}(t)$ を用いて計算される. 密度関数 $d_\beta(\boldsymbol{\omega})$ は以下の様に表される:

$$d_\beta(\boldsymbol{\omega})=\frac{\exp\left\{\beta\sum_k \ln P(\boldsymbol{a}_k|\omega_k,\boldsymbol{T})+\lambda\sum_{l>k}\left(\boldsymbol{J}(\alpha,r_{kl})\right)_{s_k s_l}\right\}}{\sum_{\boldsymbol{s}}\exp\left\{\beta\sum_k \ln P(\boldsymbol{a}_k|\omega_k,\boldsymbol{T})+\lambda\sum_{l>k}\left(\boldsymbol{J}(\alpha,r_{kl})\right)_{s_k s_l}\right\}}.$$

指標 $\boldsymbol{s}$ の和は全てのアミノ酸残基における階級化された置換速度比の n 階級に渡る和を表している. この密度関数をギブスサンプラーとして用いる.

各階級 ($s_k=1,2,3$) での ω_k の値はサンプリングの 1 ステップ毎に(4.4)式の尤度を最大にするように更新される. 図 3 (a) に $\lambda=4.0$, $\alpha=0.165$, $\beta=1.0$ の場合で, ω_k の値の更新されていく様子を 10 万ステップ示した. ω_k の値は 1 ステップ毎に更新されるが, 図 3 (a) では過去 1000 ステップ分の平均値をもってそのステップでの更新に当てている. $s_k=1$ の階級では ω_k の値

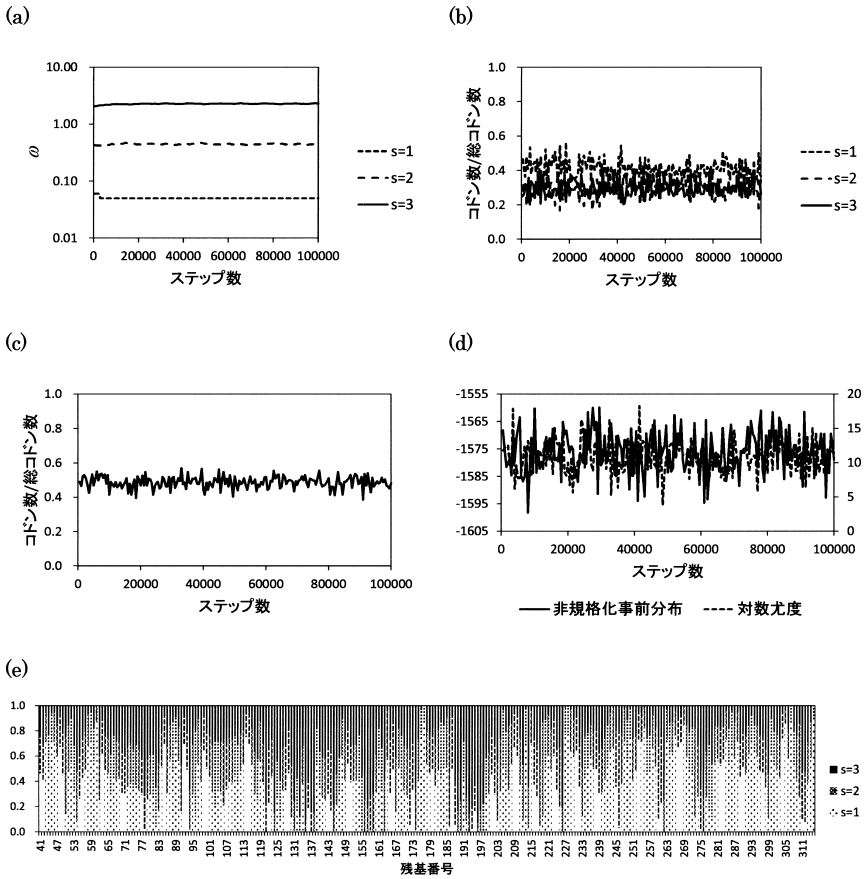

図 4. $\lambda = 4.0$, $\alpha = 0.185$, $\beta = 1.0$ の場合. (a) ω の階級値の MCMC 軌跡. (b) 3 階級の組成の MCMC 軌跡. (c) 階級遷移率の MCMC 軌跡. (d) 非規格化事前分布と対数尤度の MCMC 軌跡. (e) 各残基における ω の 3 階級の事後確率.

はおよそ 0.2 の周辺で安定しており，選択圧が変異を浄化するように働く領域である．$s_k = 2$ の階級ではおよそ 2.6，$s_k = 3$ の階級ではおよそ 16.0 の周辺で安定している．特に $s_k = 2$ の階級で $\omega_k \sim 1$ となるように制限を課していないので，いずれの階級でも選択圧が多様化をもたらす領域にある．またサンプリングの 1 ステップは，残基毎に s_k の更新を行い，これを配列全体に渡り行うことで構成している．図 3 (b) にはアミノ酸残基がどの階級にあるかをステップ毎にその割合を示した．500 ステップ毎に描画している．各ステップで約 80%の残基が変異の浄化される階級にある．図 3 (c) には階級を変更した残基が全体に占める割合を示した．約 20%の残基が各ステップでその階級を変更している．図 3 (d) には (4.3) 式と (4.4) 式の計算結果を対数で示している．図 3 (e) には各残基で ω の 3 階級の事後確率を 2 万 ~10 万ステップで平均した割合として示した．多様化圧を受ける残基が 120 番から 200 番の残基で存在することが伺える．

図 4 には $\lambda = 4.0$, $\alpha = 0.185$, $\beta = 1.0$ の場合を図 3 と同様に示した．$\alpha = 0.185$ では $\alpha = 0.165$ に比べて残基間の相関が弱くなるため，各残基独自での傾向が強く現れる結果となる．(立体構造情報を取り入れない解析に近くなる.) 各階級に属する残基の割合は各ステップでどの階級

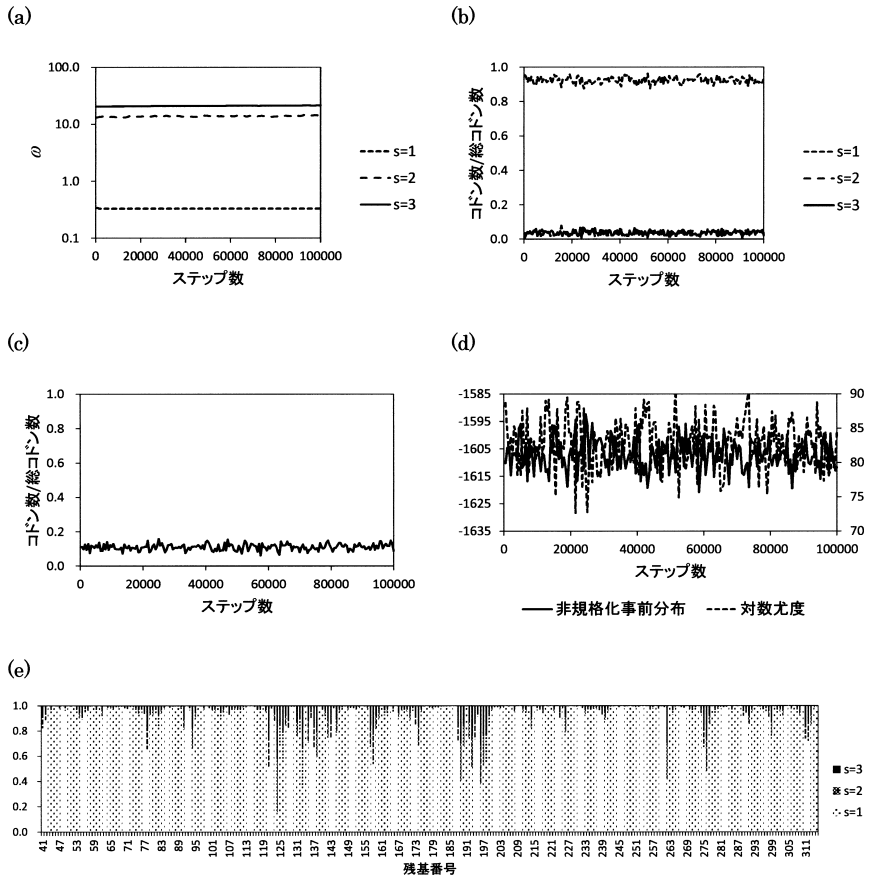

図 5. $\lambda = 4.0$, $\alpha = 0.155$, $\beta = 1.0$ の場合. (a) ω の階級値の MCMC 軌跡. (b) 3 階級の組成の MCMC 軌跡. (c) 階級遷移率の MCMC 軌跡. (d) 非規格化事前分布と対数尤度の MCMC 軌跡. (e) 各残基における ω の 3 階級の事後確率.

でも 30～40%程度であり (図 4 (b)), また約半数の残基において各ステップで階級が変更されている (図 4 (c)). これは各残基においてそれのみでの情報量が不十分であるため, 所属階級を頻繁に変更していることの現れであると考えられる. ω_k の値はいずれの階級でも $\alpha = 0.165$ に比べて低く出ている (図 4 (a)). $\alpha = 0.185$ での特徴は, 図 4 (e) に示されているようにほとんどの残基で多様化圧を受けているように出てしまっていることである. 多様化圧検出の意味では偽陽性の疑いが出てくることになる.

一方, 図 5 に示されているように $\lambda = 4.0$, $\alpha = 0.155$, $\beta = 1.0$ の場合には事情が一変する. $\alpha = 0.155$ では $\alpha = 0.165$ に比べて残基間の相関が強くなるため, 配列全体での傾向に従う結果となる. (配列全体を等しく扱った解析に近くなる.) $s_k = 1$ の階級に属する残基の割合が各ステップで 90%強あり (図 5 (b)), 残りの 10%弱が他の階級に属する. また各ステップで階級が変更されている残基は全体の 10%程度にとどまっている (図 5 (c)). これは残基間の相関が強いため, 階級の変更が抑制されているためであろう. ω_k の値はいずれの階級でも $\alpha = 0.165$ に比べて高く出ている (図 5 (a)). $\alpha = 0.155$ での特徴は, 図 5 (e) に示されているように多様化圧を受けていると結論できる残基が $\alpha = 0.165$ に比べて少ないことである.

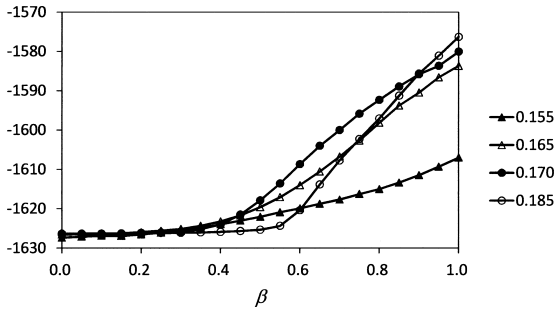

図 6. $E_\beta[\ln P]$ のサンプリング結果.

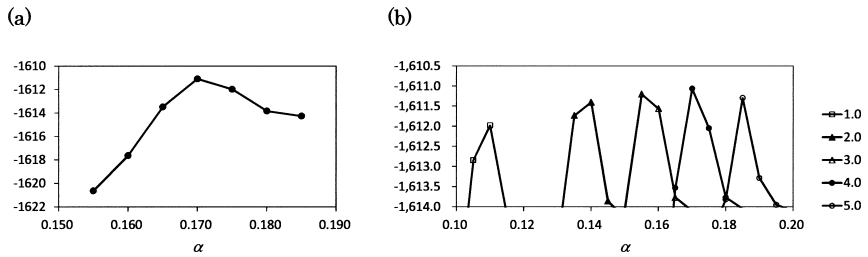

図 7. 周辺尤度の α 依存性. (a) $\lambda=4.0$ の場合. (b) $\lambda=1.0\sim5.0$ の場合 (1.0 刻み).

これまで α の値について特徴的な 3 例を観てきた. それではどの α の値を採用すればよいのであろうか. β の数値積分を遂行して (λ,α) が与えられたときの周辺尤度を計算することで, (λ,α) の値を決定する.

4.4 数値計算による熱力学的積分

熱力学的積分を数値計算によって遂行する. 導入したパラメータ β について 0 から 1 までを分割(ここでは 20 等分)し, 各 β_i $(i=0,\ldots,20)$ においてギブスサンプリングする. $\beta=0$ $(i=0)$ である場合, (4.4)式の尤度からの寄与はなく, (4.3)式の事前分布のみによる相関を観ることになる. この場合, パラメータ (λ,α) の値に関係なくクラスター構造は観られなかった. 各階級 $(s_k=1,\ 2,\ 3)$ での ω_k はほぼ同じ値をとり, 各残基でのサンプリング中の階級滞在時間は 3 つの階級で同程度であった. 図 6 に $\lambda=4.0$ の場合での各 β_i $(i=0,\ldots,20)$ における $E_\beta[\ln P]$ ((4.2)式右辺の被積分関数)のサンプリング結果を示している. $E_\beta[\ln P]$ の β 依存性はなめらかなものであると言える. これらから周辺尤度の α 依存性が得られた(図 7 (a)). この場合, $\alpha=0.17$ 周辺において最大の周辺尤度が得られている. $\lambda=1.0\sim5.0$ の場合(1.0 刻み)で同様に計算したところ, $(\lambda,\ \alpha)=(4.0,0.17)$ 周辺で最大周辺尤度を得るであろうことが判った(図 7 (b)). 図 8 はこの場合における選択圧の空間分布を示している. 受容体結合領域(インフルエンザウイルス HA タンパク質の先端に位置している)周辺で多様化をもたらす高い選択圧がかかっている様子が伺える. このように最適なモデルパラメータは熱力学的積分によって決定することが可能であり, イジング模型を用いた解析によってインフルエンザウイルス HA タンパク質における多様化を促す選択圧の空間分布を推定することが出来た.

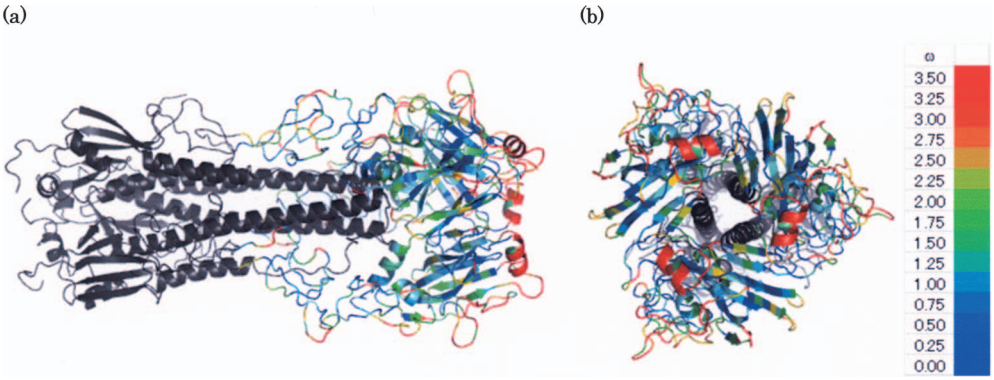

図 8. Flu-HA タンパク質上に分布する選択圧．ω の値を色で表している．低い値を青で表し，ω >3.5 を赤で表示している．解析対象でない領域は黒で表示している．立体構造データは Protein Data Bank より得た (PDB code：1HGF)．(a) HA タンパク質の側面を表示している．左側がウイルス粒子の膜側になる．(b) HA タンパク質を先端からウイルス粒子の膜へ向けて眺めている．

5. まとめ

インフルエンザウイルス HA タンパク質において多様化を促す高い選択圧が受容体結合領域周辺で得られたことは自然なことである．また過剰な高い選択圧の分布は観られず，偽陽性が押さえられているものと考えられる．我々の検出方法の妥当性を示しているものと言えるだろう．この方法を選択圧の時間分布，つまり系統樹上での分布が検出出来るように拡張することが次の課題として残っている．先にも触れたように進化系統樹のどの枝 (時間的位置) でどの空間的部位の選択圧が強く (又は弱く) なっているかを検出することが，ウイルスと宿主免疫系の関わりを深く理解することにつながる．特に幹を成す経路で選択圧が強く出るのか，末梢の枝で起こることなのかを知ることが，ウイルス独特の系統樹構造の理解を更に深めることになる．これについては検出方法を拡張し，解析結果の紹介を別の機会に出来ればよいと考えている．

参 考 文 献

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J. and Lipman, D. (2008). The influenza virus resource at the National Center for Biotechnology, *Information Journal of Virology*, **82**, 596–601.

Cox, N. J. and Bender, C. A. (1995). The molecular epidemiology of influenza viruses, *Seminars in Virology*, **6**, 359–370.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, **17**, 368–376.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Statistical Science*, **13**, 163–185.

Inoue, J. and Tanaka, K. (2001). Dynamics of the maximum marginal likelihood hyperparameter estimation in image restoration: Gradient descent versus expectation and maximization algorithm, *Physical Review E*, **65**, 016125-1–016125-11.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cam-

bridge.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration, *Systematic Biology*, **55**, 195–207.

Ogata, Y. (1989). A Monte Carlo method for high dimensional integration, *Numerische Mathematik*, **55**, 137–157.

Schlicht, R. and Iwasa, Y. (2004). Forest gap dynamics and the Ising model, *Journal of Theoretical Biology*, **230**, 65–75.

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E. and Fouchier, R. A. M. (2004). Mapping the antigenic and genetic evolution of influenza virus, *Science*, **305**, 371–376.

Suzuki, Y. (2004a). New methods for detecting positive selection at single amino acid sites, *Journal of Molecular Evolution*, **59**, 11–19.

Suzuki, Y. (2004b). Three-dimensional window analysis for detecting positive selection at structural regions of proteins, *Molecular Biology and Evolution*, **21**, 2352–2359.

Suzuki, Y. and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites, *Molecular Biology and Evolution*, **16**, 1315–1328.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation, *Journal of the American Statistical Association*, **82**, 559–567.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites, *Genetics*, **155**, 431–449.

Watabe, T. and Kishino, H. (2010). Structural considerations in the fitness landscape of a virus, *Molecular Biology and Evolution*, **27**, 1782–1791.

Watabe, T. and Kishino, H. (2012). in preparation.

Watabe, T., Kishino, H., de Oliveira Martins, L. and Kitazoe, Y. (2007). A likelihood-based index of protein-protein binding affinities with application to influenza HA escape from antibodies, *Molecular Biology and Evolution*, **24**, 1627–1638.

316 Proceedings of the Institute of Statistical Mathematics Vol. 60, No. 2, 305–316 (2012)

Spatial Distribution of Selection Pressure on a Virus Protein Deriving Its Adaptation to the Environment

Teruaki Watabe[1] and Hirohisa Kishino[2]

[1]Center of Medical Information Science, Kochi University
[2]Graduate School of Agriculture and Life Science, University of Tokyo

Proteins adapt to environments by gaining and/or obtaining functions. The adaptation to an environment is achieved by substituting the amino acid sequence and the amino acid substitution results from selection of mutations on a protein-coding gene. Hence mutations on a protein-coding gene are under the selection pressure of the environment and the strength and character of selection pressure may vary among the temporal domains in an evolutionary process. Thus, revealing the spatio-temporal fluctuation of the selection pressure improves our knowledge of adaptive evolution of the protein. We developed a method for detecting the spatial fluctuation of the selection pressure on a protein based on the hierarchical Bayesian model. The prior distribution of spatial aggregation of selection pressure is described by the Ising model, which has a theoretical framework established in the field of magnetic material physics. The hyper-parameters that define the strength and range of the spatial clustering are estimated by maximizing the marginal likelihood. The model of the prior-distribution is hard to normalize. Thus, we estimated the log marginal likelihood based on the thermodynamic integration. We applied the method to detect the spatial fluctuation of the selection pressure on the influenza hemagglutinin protein.

Key words: Molecular evolution, selection pressure, spatial distribution, hierarchical Bayesian model, Ising model.